International Journal of Innovative
Computing, Information and Control
Volume **16**, Number **3**, June **2020**

ICIC International ©2020 ISSN 1349-4198
pp. 1101–1108

MAIN TECHNIQUES OF POLICY EVALUATION AND DECISION-MAKING ANALYSIS PLATFORM

Wei Yuan[1], Yao Liu[1], Yi Huang[1] and Ruoyun Xie[2]

[1]Institute of Scientific and Technical Information of China
No. 15, Fuxing Road, Beijing 100038, P. R. China
yuanw@istic.ac.cn

[2]School of Software and Microelectronics
Peking University
No. 5, Yiheyuan Road, Beijing 100871, P. R. China

Received November 2019; revised February 2020

Abstract. *The policy text refers to documents produced in policy activities, and has always been important tools or carriers for policy evaluation. The traditional way to analyze policy texts mainly relies on people to read and compile valuable information. This method is of low informationization, efficiency and timeliness, and it is impossible to track valuable information in real time. In this article, network embedding, PMI-Entropy and other technologies are used to modify the performance of a Policy Evaluation and Decision-Making Analysis Platform; thus decisions can be evaluated and made more conveniently and precisely in the big data environment.*

Keywords: Meaningful string extraction, Network embedding, Automatic generation of briefing, Policy evaluation

1. **Introduction.** The policy text refers to documents produced in policy activities [1], and has always been important tools or carriers for policy evaluation. Real-time access to the new policy is conducive to grasping the focus of the foreign government and their policy trends, thus drawing on relevant foreign experience and promoting the development of information science in the direction of advanced technologies.

Traditional way to analyze policy mainly relies on people to read and compile valuable information,which is of low informationization, efficiency and timeliness, and it is impossible to track valuable information in real time. In the big data environment, the policy document has gained the following new characteristics: the volume of policy documents is getting larger and larger, the documents themselves are more and more complex, and more and more research demands need to be fully satisfied. Nowadays, study on specialized online policy evaluation platform is limited and some platforms have already used big data as their core. For example, Rice et al. used a data-based platform for psychiatric nurse roles and policies, but the platform can only collect supportive public opinions [2] and it does not focus on policy evaluation and decision-making. To have real-time access to policy trends and analyze them in an easier way, the Policy Evaluation and Decision-Making Analysis Platform has been put forward. This paper will discuss the key techniques and design of the Policy Evaluation and Decision-Making Analysis Platform, in order to serve scientific research managers through resource retrieval, resource analysis, and automatic briefing, and provide basic data for decision-making.

DOI: 10.24507/ijicic.16.03.1101

2. **Review.** Policy text analysis and calculation usually relies on computer to automatically analyze or label the policy text. Some of the main tasks are word analysis, sentiment analysis, and topic models. For a long time, it is mainly based on qualitative analysis, focusing on the analysis of the background, objectives, content and effects of policies. These methods are of low informationization, efficiency and timeliness. A small number of semi-quantitative and quantitative analysis also focus on simple statistical method on the basic information such as the release time of the policy text, and also the release area and quantity. At the same time, with the introduction of computational social science [3], more and more researchers have combined text analysis and computation in the field of natural language processing with traditional social science research. The calculation of policy texts, as a new way of interpreting policy texts, has gradually become a hot issue in policy research.

The mainstream methods of policy text analysis include content analysis, metrological analysis, data processing, and data mining methods. 1) Analysis of Content. This method is a semi-quantitative research method. It mainly follows the theory of political discourse in political science and hermeneutics and the interpretation of symbolic and political words in the framework of comparative research tradition. 2) Metrological Analysis. This method mainly adopts the basic theories and methods of textual measurement analysis, and analyzes the existing policy text database or policy text corpus in terms of policy topic distribution, policy release time series distribution, policy citation and policy subject relationship [4]. There are three main types of methods and tools for the measurement and analysis of policy texts: First is the textual analysis method and tools of the policy text database, such as LexisNexis, ProQuest and other policy or legal text databases, which are combined with the field settings of the database; Second, network analysis and alternative metrology methods are used. For example, Altmetric can track information of Weibo [5]; Third, study through policy text collection and corpus construction are used to propose new statistical caliber and research methods. For example, Wilson et al. of Carnegie Mellon University conducted thematic analysis of the website privacy policy [6]. 3) Data Processing Methods. In the process of data processing, policy texts or corpus are generally applicable to general natural language processing methods and text data processing methods, including vocabulary construction, word segmentation, part-of-speech tagging, synonym merging, stop word list, feature word extraction, word frequency calculating, word frequency distribution, co-word analysis. 4) Data Mining Methods. Compared with semantic analysis, data mining methods of policy texts pay more attention to cluster features, discover related knowledge or rules in a large number of text data sets, and focus on deep potential semantic knowledge discovery. Therefore, typical methods such as policy sentiment analysis, policy opinion analysis, and government behavior prediction have received extensive attention in the field of policy research. Government behavior predictions reflect the methods and ideas of policy prediction analysis, through campaign guidelines or key policies for government leaders and political parties. In all, the existing studies generally lack systemicity and continuity.

Network embedding, also known as graph representation learning, is one of the core techniques of this research. According to the difference of the representation granularity, the graph embedding can be divided into four methods: node representation, edge representation, subgraph representation and graph representation. The network structure processed in this study is accurate to the node representation, which assumes that nodes with similar topological structures have similar vector representations. Network embedding can not only better study and analyze the relationship between nodes in complex networks, but also integrate topological structures and external complex information such as texts owned by nodes to form node representations with more classification features.

According to the content of the node, the existing network embedding algorithm can be divided into method based on network structure and method combining external knowledge. In this paper, the network embedding is used to map the multi-faceted knowledge network node information into the node representation, thus preparing for the technology briefing generation task.

3. **Design and Ideas.** The two main tasks of the Policy Evaluation and Decision-Making Analysis Platform are the construction of text processing platform and the automatic briefing and recommendation system. The key techniques of the platform include information acquisition, text analysis and other natural language processing methods for special use like network embedding, mutual information combined with information entropy (PMI method) for obtaining candidate meaningful strings, which are talked about in Section 4. After that, rules are used. The automatic briefing and recommendation system can conduct the visual analysis of the policy and achieve personalized recommendation. Policy information can be fetched from the Internet, summarized automatically and later translated to Chinese to provide real-time information for policy decision-making, analysis and research. In general, the platform is divided into six modules based on their functions: Acquisition Module, Processing Module, Retrieval Module, Analysis Module, Calculation Module and Briefing Presentation Module. These six modules start from resource storage, analysis and finally realize the presentation of policy briefing. In order to control the whole system, the System Management Module is designed, the function including user management, role management, authority management, etc.

The main function of Acquisition Module is automatically obtaining policy information by using web crawler to satisfy different needs of users. During our research, more than 20 websites and databases have been fetched, covering China, the United States, and the United Kingdom, the EU, Japan and other countries, which are prepared for briefing generation. Processing Module further analyzes and processes the acquired resources. Users can upload policy text in different formats. After that, the structure and key information of the text will be extracted and presented to the user. Retrieval Module mainly provides two methods for users: keyword retrieval and advanced retrieval which can check for both of original resources and regenerative resources such as briefing generated in the platform. The Analysis and Calculation Module mainly conducts resource visualization analysis and multi-indicator combination statistics, and provides single analysis, macro analysis, comparative analysis, policy verification analysis, and time series analysis. The visual display method is more diversified and can be used as an independent research tool or used for briefing generation.

In the Briefing Module, the briefing is generated through deep learning algorithm, which makes the text structured through the entity relationship extraction, and thus the knowledge network structure is built. The network embedding is used to represent the concept nodes and chapter nodes in the form of network structure, and is applied to generate the structure and content of the technology briefing. The customized briefing can be set by users themselves to design a template for the briefing. The system will extract the key content and support the generation of diversified statistics and automatic analysis.

4. **Advantages and Features.** The Policy Evaluation and Decision-Making Analysis Platform has carried out in-depth research and experiment on multiple technical points, and formed its unique resource processing, analysis and text generation method to ensure the policy text analysis effect and platform practicality.

4.1. **Meaningful string extraction.** The first step of text analysis is a series of text preprocessing tasks such as word segmentation, part of speech analysis, de-stopping words, and synonymous merging. For the semantic analysis of policy texts, the effect of text segmentation directly determines the accuracy of subsequent analysis. In the field of mid-word segmentation, the maximum entropy model has been widely used. Jiang used the maximum entropy model based on meaningful strings for word segmentation and achieved good performance compared with maximum entropy model only [7].

Different from other texts, the policy text tends to have more long strings which are meaningful, and also new word, also the grammatical structure is more complicated. The conventional word segmentation method cannot better preserve the semantic information of the original document. According to the research of Jiang and the unique characteristics of policy texts, the study of the system uses meaningful string extraction method that combines information entropy, mutual information with rules to make sure the performance of extraction. Mutual information combined with information entropy algorithm (PMI method) is used to obtain candidate meaningful strings. After that, rules are used to choose the most important strings from candidate meaningful strings.

In the test, 100 policy texts are randomly chosen as test corpus. Those texts are chosen from government work report in different cities in 2018. Firstly, jieba word segmentation tool is used to perform initial segmentation of text. Then meaningful strings are extracted using the above method. Corresponding to the same corpus, the extracted candidate meaningful strings are as shown in Figure 1. The Chinese characters are those meaning strings, and the figures show their scores when the threshold value is 0.1.

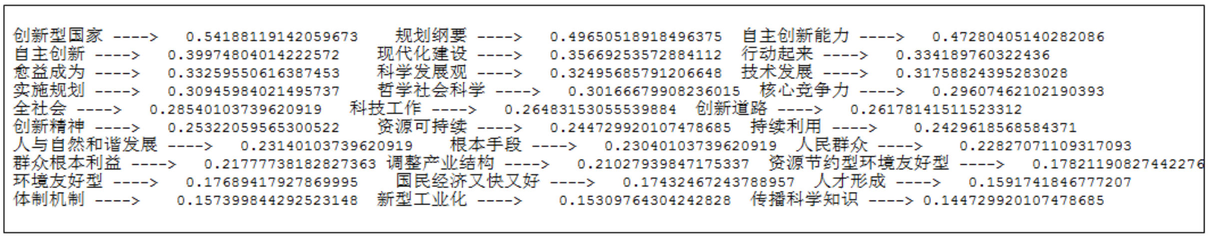

Figure 1. Result of meaningful string extraction

In order to better prove the effectiveness of the method used in the system, three methods are tested for comparison. This paper evaluates the effect of the extraction method by the accuracy rate, recall rate and F1 value. The meaning and calculation method of the evaluation indicators are as follows.

Precision: The ratio of the number of correctly extracted values to the total amount of vocabulary extracted for a meaningful string in a given text.

$$P(\text{Precision}) = \frac{TP}{TP + FP} \tag{1}$$

Recall: The ratio of the number of correctly extracted numbers to the number of meaningful strings actually determined for a meaningful string in a given text.

$$R(\text{Recall}) = \frac{TP}{TP + FN} \tag{2}$$

F score: The sum of the accuracy and recall.

$$\text{F1} = \frac{2PR}{P + R} \tag{3}$$

While TP refers to the number of correctly extracted words, FP refers to the number of vocabulary errors extracted, and FN refers to meaningful strings that are not extracted.

TABLE 1. Comparison of three different methods

Method	Precision	Recall	F1
PMI-Entropy	0.7152	0.7654	0.7394
Rules-based	0.8113	0.7043	0.7540
Rules-PMI-Entropy	0.8892	0.7468	0.8118

The experimental results are presented in Table 1.

The first two methods are used respectively to compare with the method which combines them as a whole. It can be seen from the experimental results that the recall rate is higher with the PMI-Entropy method, but the accuracy is lower because many non-meaningful strings that do not conform to the semantic expression of the policy domain are extracted in the extraction, and the rules are simply used. The extraction method, although accurate, does not cover all meaningful strings. The method used by the system can effectively make up for the deficiency of the two methods, and its accuracy and recall rate are relatively high. The method is applied to the platform, and a good keyword extraction effect is obtained. Figure 2 shows the Chinese meaningful strings which have the highest frequency, and their frequency respectively, for example, the first one in the left is listed company, which has the longest meaning.

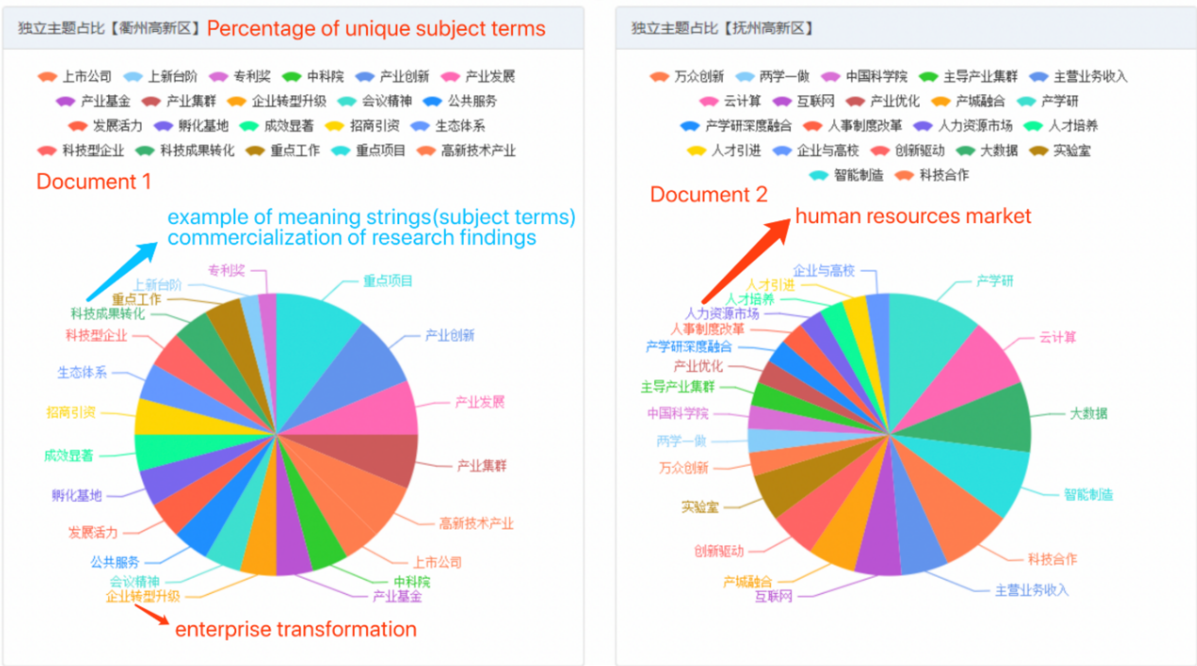

FIGURE 2. Meaningful strings extraction presentation in the platform

4.2. **Automatic generation of briefings based on network embedding.** Network embedding, also known as graph representation learning, is one of the core techniques of the platform. A significant amount of research effort is made in the past few years to generate node representations from graph-structured data using representation learning methods [8]. The existing text generation algorithms are mostly based on characters or words, ignoring the relationship between concepts. Therefore, this paper uses network embedding algorithm that can fuse the concept semantics of the knowledge network, the topology of the concept links, and the concept category labels. Combined with the knowledge reasoning model, the node representation of the concept is calculated, thus discovering the potential knowledge network information in the original text.

Through the method of network embedding, briefings can be automatically generated. The basic idea is to structure the policy text content through the entity relationship extraction method, thereby constructing the knowledge network structure, and using the network embedding method to represent the concept nodes and chapter nodes in the network structure. The concept network is finally used in the generation of the briefings.

After the knowledge network is built, method to represent the network nodes is a challenging task. The system will use the TriDNR [9] algorithm proposed by Pan et al. to make the concept nodes contain the network topology information, semantic information, label information and reasoning information. After the concept node representation is obtained, the chapter node representation can then be obtained. The chapter is in a tree structure, the leaves of the tree structure are connected with the concept nodes, and the basic method is the method of adding concepts to obtain the representation of the leaf nodes of the chapter. It can be found that the chapter nodes are equivalent to the paragraph nodes, so Doc2vec method can be adopted. Doc2vec is used to place the important text and chapter nodes in the front position. The TextRank algorithm, a key sentence extraction algorithm, is used to represent the concept node network representation as the algorithm input. The Chinese words at the right are subtitles in original corpus.

```
0.042076005607176681 2. 愿景：工业4.0作为智能、网络化世界的一部分
0.031228206896693222 2.1 塑造工业4.0愿景
0.031204141470405693 2.6 工业4.0之路
0.030919204916247241 3. 双重战略：成为领先的市场和供应商
0.030919204916247241 5.4 安保是工业4.0成功至关重要的因素
0.030652227184829544 5.3 为工业提供一个全面宽敞的基础设施
0.030652227184829543 3.3.3 纵向集成和网络化制造系统
0.030519294657362788 2.2 在工业4.0下未来会是什么样子
```

FIGURE 3. Sequence of nodes

The node vector representation of each document with text structure is used as the input of the text generation model, and the structure and content in the technology briefing are respectively generated. Through the above method, the user inputs the search term, and the platform will return the policy text corpus. When multiple articles are chosen for extraction, it is not necessary to sort the importance nodes of the search results, and directly generate the technical briefings from the selected multiple articles. The left side is the generation structure of the technology briefing, and the right side is the content of each chapter. Click on each chapter subtitle, and the chapter content is now in the text box. Users can modify it directly in the text box.

Since the technical briefing of this paper is generated based on the original science and technology policy or the knowledge network of the scientific report, this paper judges the performance of the method through coverage rate of the conceptual relationship network. That is, if the knowledge network in the generated technology briefing completely coincides with the standard knowledge network, the score is the number of concepts, and the coverage rate is 100%; If the concepts overlap but the relationship is incorrect, each concept is scored 0.5, and the final coverage is the ratio of the score to the highest score.

As seen in Table 2, method in this paper has the highest coverage rate. The traditional algorithm calculates the similarity between sentences and only considers the co-occurrence of words. Differently, this method calculates the important nodes of the concept and the important nodes of the text in the knowledge network, thus discovering the potential semantic association between the sentences, supplementing the sentences between the

FIGURE 4. Briefings presented in the platform

TABLE 2. Comparison of TextRank, TextRank+Glove, and method in this paper

Method	Coverage rate
TextRank	0.133
TextRank+Glove	0.189
Method in this paper	0.224

concepts. Semantics, which increases the weight of sentences in nodes with higher similarity, can finally extract text content that covers important information.

5. **Conclusion.** This paper studies key technologies of the Policy Evaluation and Decision-making Analysis Platform to provide reference for policy decision-makers. Through the above research, a new meaningful string method is put forward to make up the shortage of common word segmentation tools and achieve good performance in policy evaluation. What is more, network embedding is used in generating briefing automatically, which has good performance when it is integrated in our platform. In future study, accurate and standard resource acquisition should be prepared to ensure that the content of the briefing is of a higher level, and more subjects of briefing can be generated to meet users' need.

Acknowledgment. This work is partially supported by ISTIC Key Project Program (Grant No.: ZD2018-11). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] L. Geng, J. Sun and Z. Zhou, Policy text calculation: A new way of interpreting policy texts, *Library and Information*, no.6, pp.47-55, 2016.
[2] M. J. Rice, J. Stalling and A. Monasterio, Psychiatric-mental health nursing: Data-driven policy platform for a psychiatric mental health care workforce, *Journal of the American Psychiatric Nurses Association*, vol.25, no.1, pp.27-37, 2019.

[3] D. Lazer, A. Pentland, L. Adamic et al., Computational social science, *Science*, vol.323, no.1, pp.721-723, 2009.

[4] J. Li, Y. Liu, C. Huang et al., Analysis of the reconstruction policy text data by bibliometric research – The origin, migration and method innovation of policy bibliometrics, *Journal of Public Administration*, no.2, pp.138-144, 2015.

[5] H. Piwowar, Altmetrics: Value all research products, *Nature*, vol.493, p.159, DOI: https://doi.org/10.1038/493159a, 2013.

[6] S. Wilson, F. Schaub, R. Ramanath et al., Crowdsourcing annotations for websites' privacy policies: Can it really work?, *Proc. of the 25th International Conference on World Wide Web*, pp.133-143, 2016.

[7] M. Jiang, *Maximum Entropy Chinese Word Segmentation Method Based on Effective Substrings*, Master Thesis, Tianjin University of Finance and Economics, 2018.

[8] A. Mohan and K. V. Pramod, Network representation learning: Models, methods and applications, *SN Applied Sciences*, vol.1, Article No. 1014, 2019.

[9] S. Pan, J. Wu, X. Zhu et al., Tri-party deep network representation, *Network*, vol.11, no.9, 2016.